Published online 26 April 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W413–W416***
doi:10.1093/nar/gkp281

SITEHOUND-web: a server for ligand binding site identification in protein structures

Marylens Hernandez, Dario Ghersi and Roberto Sanchez*

Department of Structural and Chemical Biology, Mount Sinai School of Medicine, New York, NY 10029, USA

Received March 1, 2009; Revised April 6, 2009; Accepted April 14, 2009

ABSTRACT

SITEHOUND-web (http://sitehound.sanchezlab.org) is a binding-site identification server powered by the SITEHOUND program. Given a protein structure in PDB format SITEHOUND-web will identify regions of the protein characterized by favorable interactions with a probe molecule. These regions correspond to putative ligand binding sites. Depending on the probe used in the calculation, sites with preference for different ligands will be identified. Currently, a carbon probe for identification of binding sites for drug-like molecules, and a phosphate probe for phosphorylated ligands (ATP, phoshopeptides, etc.) have been implemented. SITEHOUND-web will display the results in HTML pages including an interactive 3D representation of the protein structure and the putative sites using the Jmol java applet. Various downloadable data files are also provided for offline data analysis.

INTRODUCTION

The combination of Structural Genomics efforts and computational modeling has resulted in a large amount of 3D structure information for proteins. However, to a large degree, this structural information has not been translated into functional information. For example, understanding substrate specificity, catalysis or inhibition, is still largely dependent on biochemical and biophysical analysis of individual proteins. While protein structure in principle encodes this mechanistic information, reliable computational tools and approaches to establish a connection between structure and function are still lacking. The molecular function of proteins is largely determined by their interaction with other molecules at binding sites on the protein surface. Thus, localization and characterization of a ligand binding site can contribute to functional annotation of a protein; it can guide mutational experiments, and be useful in predicting or verifying interactions. The identification of ligand binding sites can also be an important part of the drug discovery process. Knowing the location of binding sites facilitates virtual screening for hits, lead optimization and identification of features that influence the selectivity of binding. Hence, several methods have been developed for the identification of binding sites from protein structures (1–6) and sequences (7–10). The structure-based methods recognize geometrical features, such as clefts, or energetic features that describe the molecular interaction properties of the protein surface. In general, structure-based methods can be seen as complementary to sequence-based methods that exploit evolutionary information. Here, we describe the SITEHOUND-web server for identification of ligand binding sites in protein structures. It uses an energy-based approach to identify regions with high potential for interaction with ligands. A unique feature of SITEHOUND-web is that it implements the use of different probes to characterize a protein structure, which enables not only the identification of different types of binding sites, but also a preliminary description of its interaction properties.

METHODS

The SITEHOUND algorithm

The SITEHOUND algorithm identifies potential ligand binding sites by recognizing regions characterized by favorable non-bonded interactions with a chemical probe (6). Depending on the nature of the probe, different types of binding sites can be identified. Currently, a 'Carbon' probe and a 'Phosphate' probe are available for the identification of binding sites for drug-like molecules, and ligands containing phosphate groups, respectively. Affinity Maps (also called Molecular Interaction Fields) describing the interaction of the probe and the protein on a regular 3D lattice are calculated using either the AutoGrid program (11) for the Carbon probe, or the EasyMIFs program (D. Ghersi and R. Sanchez, manuscript submitted for publication) for the Phosphate probe. SITEHOUND then filters the affinity map points corresponding to unfavorable interaction energies. The remaining points are clustered according to their spatial proximity using an agglomerative hierarchical

*To whom correspondence should be addressed. Tel: +1 212 659 8648; Fax: +1 212 659 8232; Email: roberto@sanchezlab.org or roberto.sanchez@mssm.edu

clustering algorithm. The final output is a list of 'interaction energy clusters' corresponding to putative binding sites, which are ranked by Total Interaction Energy (TIE) (the sum of the energy values of all the points that belong to the same cluster). A test study carried out on 77 experimentally determined protein structures, corresponding to known protein–ligand complexes, showed that the correct binding site was among the top three SITEHOUND clusters in 95% of the cases (6).

Server implementation

SITEHOUND-web (http://sitehound.sanchezlab.org) was implemented using a python-CGI and JavaScript based platform. A series of python 'wrappers' integrate programs MODELLER (12), AutoGrid (11), EasyMIFs (D. Ghersi and R. Sanchez, manuscript submitted for publication), and SITEHOUND (6), resulting in a completely automated identification of ligand binding sites from a standard PDB file. The input PDB file is first scanned for ligands and chain composition using MODELLER. Any existing ligands are removed to avoid interference with binding site identification. The processed PDB file is then passed to either AutoGrid or EasyMIFs, depending on the user-selected probe. The resulting affinity map is then passed to SITEHOUND. The output is displayed using HTML pages including an interactive 3D representation of the protein structure and the putative binding sites using the Jmol java applet (http://www.jmol.org).

SITEHOUND-web input

SITEHOUND-web requires a PDB file as input and the specification of a probe and clustering algorithm for the calculation. The input PDB file can either be uploaded or a PDB code can be specified. When specifying a PDB code the corresponding file is copied from the PDB database. The PDB file does not need to be preprocessed (e.g. removal of ligands) since the server does this automatically. Two types of probes are currently available: a carbon probe for the identification of binding sites for molecules that interact mainly through van der Waals contacts; and a phosphate probe which is used to identify sites that bind to phosphorylated ligands. The carbon probe has been validated mainly with drug-like molecules (6) and the phosphate probe with phosphopeptides, phosphosugars, and ATP (D. Ghersi and R. Sanchez, manuscript in preparation). Finally, a clustering algorithm needs to be selected. The clustering algorithm determines the way in which SITEHOUND combines individual affinity map points into clusters corresponding to putative binding sites. The average-linkage clustering tends to result in relatively spherical clusters and is the default for both probes. While only the use of average-linkage clustering has been tested extensively in SITEHOUND, the single-linkage clustering algorithm is provided as an alternative to be used with the carbon probe for the identification of larger elongated binding sites, like those of peptides. The SITEHOUND-web input page also provides sample input files and output data. Once a request has been submitted, the calculation proceeds unless a multiple chain PDB file has been uploaded or selected. In this case, the server will provide the option to select one or more chains from the PDB file to be included in the calculation. After chain selection the calculation proceeds. For a medium-sized protein (150 residues), a typical calculation takes ~1 min. However, running time also depends on the shape of the protein, with elongated proteins taking longer than spherical ones.

SITEHOUND-web output

The output of SITEHOUND-web has two components: an interactive web screen displaying a summary of results with a 3D representation of the putative binding sites on the protein structure; and downloadable files for offline analysis.

The output screen is divided into five sections (Figure 1). A 'Cluster Data' table (Figure 1A) displays the top 10 ranking interaction energy clusters (i.e. putative binding sites). This table shows the rank, TIE, coordinates, and volume for each cluster. The color of the rank corresponds to the color of the cluster in the 3D display. The TIE, which is used to rank the clusters, is an indication of the strength of the clusters. Significant clusters usually have TIEs that stand out against the background of weaker clusters (see clusters 1 and 2 in Figure 1A; and cluster 1 in Figure 2A). The cluster coordinates correspond to the x, y and z coordinates of the center of each cluster. This can be used, for example, to set up a docking box centered around a putative binding site (6). Finally, the volume of the cluster in $Å^3$ is displayed in the last column. A 3D interactive view of the protein structure and the clusters (Figure 1B) is provided using the Jmol molecular viewer. This view interacts with the 'Cluster Selection' panel (Figure 1C), which can be used to toggle the display of any of the top 10 clusters on and off. The coloring of the clusters corresponds to their rank in the Cluster Data table. A 'Cluster Details' panel (Figure 1D) provides a list of protein residues in the vicinity of a selected cluster. Clicking on its corresponding rank in the Cluster Data table changes the selected cluster. Finally, the 'Download Data' panel (Figure 1E) provides links to various data files. The 'Cluster Data' file provides the same information as the Cluster Data table, but for all identified clusters. The DX file stores cluster data in the DX format which is useful for display in programs such as PyMOL (http://www.pymol.org) and Chimera (13). The Cluster PDB file contains the coordinates of the cluster points in PDB format; it can be used to display the clusters in most molecular viewers (Figure 3) and is the file used internally by SITEHOUND-web to display the clusters using Jmol. The MAP file is the affinity map used for the identification of binding sites. It can be used with the offline version of SITEHOUND (D. Ghersi and R. Sanchez, manuscript submitted for publication) to explore different parameters for cluster analysis.

CONCLUSIONS

Ligand binding site identification is an important tool in structural biology because it can bridge the structure-

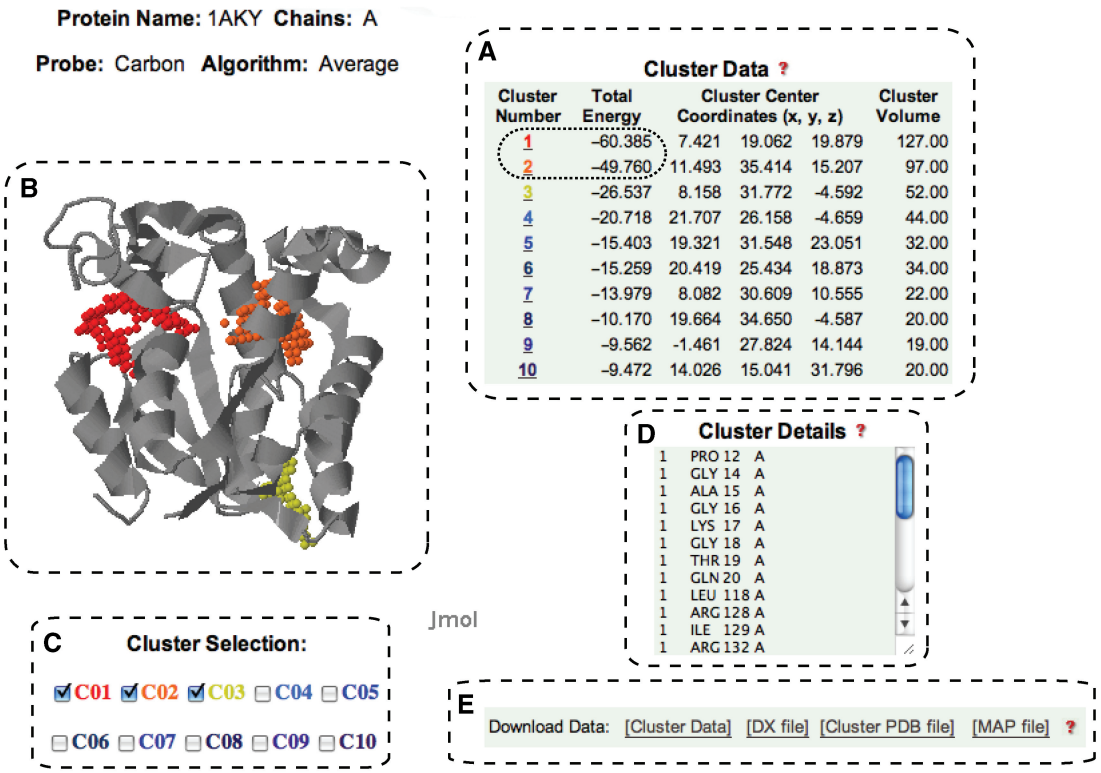

Figure 1. SITEHOUND-web Carbon probe output example. The output for yeast adenylate kinase (14) (PDB code 1aky) processed with the carbon probe and the average-linkage clustering algorithm is shown. (**A**) The 'Cluster Data' table summarizes the information for the top 10 clusters ranked by Total Interaction Energy. The Cluster Number indicates the rank of the cluster with the colors corresponding to the coloring of the cluster in the structure display and cluster selection windows. Two clusters (circled with the dotted line) stand out has having significantly more favorable interaction energy than the rest. The coordinates for the center of the cluster and the cluster volume are also displayed. (**B**) The structure display window provides a 3D view of the clusters in the context of the protein structure using the Jmol java applet (http://www.jmol.org). Up to 10 clusters can be displayed. (**C**) The 'Cluster Selection' panel allows toggling the display of individual clusters on or off. By default, the top-three ranking clusters are selected. (**D**) The 'Cluster Details' panel displays all residues in contact with the cluster selected in the Cluster Data window. (**E**) The 'Download Data' panel provides links to various data files for offline analysis (see text for a description of each file).

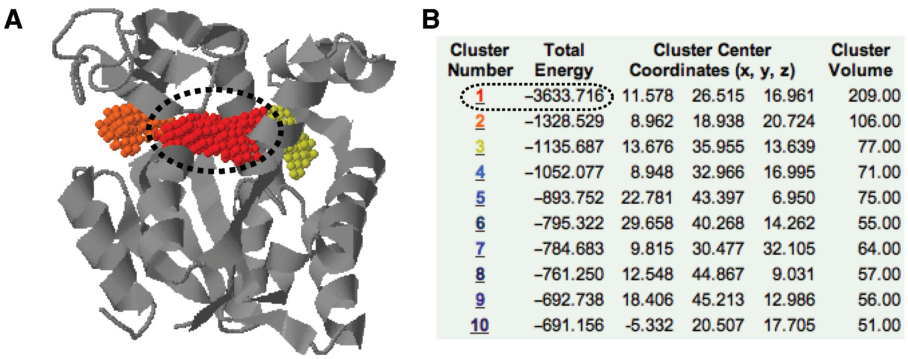

Figure 2. SITEHOUND-web Phosphate probe output example. The output for yeast adenylate kinase (14) (PDB code 1aky) processed with the phosphate probe and the average-linkage clustering algorithm is shown. Only the Structure Display (**A**) and Cluster Data (**B**) panels are shown. Cluster 1 (circled) stands out as having significantly more favorable interaction energy with the phosphate probe than the rest of the clusters. The position of cluster 1 is intermediate between the two most favorable Carbon probe clusters (Figures 1 and 3).

function gap in a homology-independent way. SITEHOUND-web is a ligand binding site identification server that can provide information about the location and binding preference of sites in protein structures. It has a simple interface that only requires the user to select a protein structure and two options (probe and clustering algorithm). A unique feature of SITEHOUND-web is its ability to identify different types of binding sites depending on the probe used for calculation. Future development of SITEHOUND will include the addition of more probes for characterization of a more diverse set of sites. Because the method requires

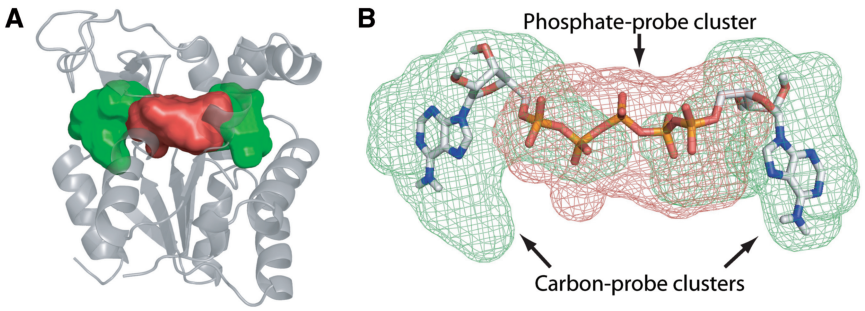

Figure 3. Combining SITEHOUND-web outputs to describe the adenylate kinase binding site. (**A**) Ribbon diagram of the yeast adenylate kinase structure showing the top ranking clusters from Figures 1 and 2 as solid surfaces: phosphate probe cluster (red) and carbon probe clusters (green). (**B**) SITEHOUND-web clusters superposed on the structure of the Ap5A (bis(adenosine)-5′-pentaphosphate) inhibitor of adenylate kinase (14). The phosphate probe correctly identifies the pathway of phosphoryl transfer, and the carbon probe correctly identifies the adenosine binding regions. The figure was prepared using the 'Cluster PDB file' downloadable output files from SITEHOUND-web examples shown in Figures 1 and 2, and the PyMOL molecular graphics program (http://www.pymol.org).

only the structure of the protein as input it can be used to complement sequence-based methods for identification of functionally important residues, which rely on evolutionary information. We expect SITEHOUND-web to be especially useful in the context of structural annotation, and docking applications in which the binding site is unknown. While binding site identification methods can help in locating and characterizing the regions of the protein to which a ligand may bind, they cannot guarantee that a given site will or will not bind a ligand. This is a problem that is better addressed by techniques such as virtual screening that can be carried out on the putative binding sites.

ACKNOWLEDGEMENTS

The authors thank the members of the Sanchez lab for useful suggestions and discussions. D.G. is a student in the New York University/Mount Sinai Computational Biology IGERT program. R.S. is an Irma T. Hirschl Career Award recipient.

FUNDING

National Science Foundation (MCB 0517352); the National Institutes of Health (GM081713). Funding for open access charge: National Institutes of Health (GM081713).

Conflict of interest statement. None declared.

REFERENCES

1. Bartlett,G.J., Todd,A.E. and Thornton,J.M. (2003) Inferring protein function from structure. *Methods Biochem. Anal.*, **44**, 387–407.
2. Campbell,S.J., Gold,N.D., Jackson,R.M. and Westhead,D.R. (2003) Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol.*, **13**, 389–395.
3. Laskowski,R.A., Thornton,J.M., Humblet,C. and Singh,J. (1996) X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins. *J. Mol. Biol.*, **259**, 175–201.
4. Laurie,A.T. and Jackson,R.M. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, **21**, 1908–1916.
5. Liang,S., Zhang,C., Liu,S. and Zhou,Y. (2006) Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res.*, **34**, 3698–3707.
6. Ghersi,D. and Sanchez,R. (2009) Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites. *Proteins*, **74**, 417–424.
7. Lichtarge,O. and Sowa,M.E. (2002) Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.*, **12**, 21–27.
8. Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.
9. Berezin,C., Glaser,F., Rosenberg,J., Paz,I., Pupko,T., Fariselli,P., Casadio,R. and Ben-Tal,N. (2004) ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics*, **20**, 1322–1324.
10. del Sol Mesa,A., Pazos,F. and Valencia,A. (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, **326**, 1289–1302.
11. Morris,G.M., Goodsell,D.S., Halliday,R.S., Huey,R., Hart,W.E., Belew,R.K. and Olson,A.J. (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, **19**, 1639–1662.
12. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
13. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
14. Abele,U. and Schulz,G.E. (1995) High-resolution structures of adenylate kinase from yeast ligated with inhibitor Ap5A, showing the pathway of phosphoryl transfer. *Protein Sci.*, **4**, 1262–1271.